

# Reconstructing Undirected Graphs from Eigenspaces

**Yohann De Castro**

YOHANN.DECASTRO@MATH.U-PSUD.FR

*Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay,  
F-91405 Orsay, France*

**Thibault Espinasse**

ESPINASSE@MATH.UNIV-LYON1.FR

*Institut Camille Jordan (CNRS UMR 5208), Université Claude Bernard Lyon 1,  
F-69622 Villeurbanne, France*

**Paul Rochet**

PAUL.ROCHET@UNIV-NANTES.FR

*Laboratoire de Mathématiques Jean Leray (CNRS UMR 6629), Université de Nantes,  
F-44322 Nantes, France*

**Editor:** ArXiv:1603.08113v2

## Abstract

In this paper, we aim at recovering an undirected weighted graph of  $N$  vertices from the knowledge of a perturbed version of the eigenspaces of its adjacency matrix  $W$ . Our approach is based on minimizing a cost function given by the Frobenius norm of the commutator  $AB - BA$  between symmetric matrices  $A$  and  $B$ .

In the Erdős-Rényi model with no self-loops, we show that identifiability (*i.e.*, the ability to reconstruct  $W$  from the knowledge of its eigenspaces) follows a sharp phase transition on the expected number of edges with threshold function  $N \log N/2$ .

Given an estimation of the eigenspaces based on a  $n$ -sample, we provide support selection procedures from theoretical and practical point of views. In particular, when deleting an edge from the active support, our study unveils that our test statistic is the order of  $\mathcal{O}(1/n)$  when we overestimate the true support and lower bounded by a positive constant when the estimated support is smaller than the true support. This feature leads to a powerful practical support estimation procedure when properly thresholding. Simulated and real life numerical experiments assert our new methodology.

**Keywords:** Support selection; Backward algorithm; Identifiability; Graph; Eigenspaces;

## 1. Presentation

We investigate the reconstruction of an undirected weighted graph of size  $N$  from incomplete information on its set of edges (for instance, one knows that the target graph has no self-loops) and an estimation of the eigenspaces of its adjacency matrix  $W$ . This situation depicts any model where one knows in advance a linear operator  $K$  that commutes with  $W$ . Several examples are presented in Section 3 while the general model is given in Section 2.1.

Section 2.2 is concerned with identifiability issues, *i.e.* the capacity to solve such problem. We exhibit sufficient and necessary conditions on the ability to reconstruct an undirected graph with no self-loops from the knowledge of the eigenspaces of  $W$ . These conditions allow us to derive a sharp phase transition on identifiability in the Erdős-Rényi model.

Then, we introduce and theoretically assert new estimation schemes based on the Frobenius norm of the commutator  $AB - BA$  between symmetric matrices  $A$  and  $B$ , see Section 4.1. More precisely, we assume that we have access to an estimation  $\hat{K}$  of  $K$  build from a  $n$ -sample and we consider the empirical contrast given by the commutator, namely  $A \mapsto \|\hat{K}A - A\hat{K}\|$  where  $\|\cdot\|$  denotes the Frobenius norm. Using backward-type procedures based on this empirical contrast, Section 4 derives estimators of the graph structure, *i.e.*, its set of edges  $S^*$  (referred to as

the support). This studies reveals typical behaviors of the empirical contrast when the estimated support  $S$  (referred to as the active set) contains or not the true support  $S^*$ . Numerical experiments developed in Section 5 (simulated data) and Section 6 (real life data) assess the performances of our new estimation method. Discussion and related questions are presented in Section 7.

To the best of our knowledge, the framework of this paper is new and the present results solve the identifiability issues and enforce an efficient backward-type estimation procedure. Related topics encompass spectral, least-squares and moment methods for graph reconstruction Verzelen et al. (2015); Guédon and Vershynin (2015); Klopp et al. (2015); Bubeck et al. (2016), Graphical Models Verzelen (2008); Giraud et al. (2012), or Vectorial AutoRegressive process Hyvärinen et al. (2010) to name but a few. In the specific cases of Ornstein-Uhlenbeck processes and non-linear diffusions, the interesting papers Bento et al. (2010) and Bento and Ibrahimi (2014) tackle a related subproblem that is to estimate  $W$  along a trajectory, see Section 3.6 for further details. Note that the framework of the present paper addresses processes observed at i.i.d. random times (with possibly unknown distribution) which are not cover by Bento et al. (2010) and Bento and Ibrahimi (2014).

## 2. Model and Identifiability

### 2.1 The Model

Consider a symmetric matrix  $W \in \mathbb{R}^{N \times N}$  with some zero entries, where nonzero entries describe the intensity of a link of any form of local interaction. One may understand  $W$  as the adjacency matrix of an undirected weighted graph with  $N$  vertices. We focus on the eigenspaces of  $W$  examining models where we have no information on the spectrum of the graph. Depicting this situation, we assume that the information on the target  $W$  stems from an unknown transformation  $K = f(W) \in \mathbb{R}^{N \times N}$  or, in more realistic scenarios, from a perturbed version  $\hat{K}$  of  $K$ . Here,  $f$  is assumed to be an injective analytical function on the real line so that the transformation  $K = f(W)$  may be understood as an operation on the spectrum of  $W$  only, stabilizing the eigenspaces. Therefore,  $W$  and  $K$  share the same eigenspaces and in particular, they commute, *i.e.*,  $WK = KW$ .

Our goal is to uncover  $W$  from the knowledge of an estimator  $\hat{K}$  of  $K$ , namely reconstruct  $W$  from a perturbed observation of its eigenspaces. The key point is then to use extra information given by the location of some zero entries of  $W$ . Hence, we assume that one knows in advance a set  $F \subset [1, N]^2$  of “forbidden” entries such that

$$\forall (i, j) \in F, \quad W_{ij} = 0 \quad (\mathbf{H}_F)$$

Equivalently, the set  $F$  is disjoint to the set of edges of the target graph. Throughout this paper, a special case of interest is given by  $F = F_{\text{diag}} := \{(i, i) : 1 \leq i \leq N\}$  conveying that there are no self-loops in  $W$ .

### 2.2 Identifiability

For  $S \subseteq [1, N]^2$ , denote by  $\mathcal{E}(S)$  the set of symmetric matrices  $A$  whose support is included in  $S$ , which we write  $\text{Supp}(A) \subseteq S$ . Given the set  $F$  of forbidden entries defined via  $(\mathbf{H}_F)$ , the matrix of interest  $W$  is sought in the set  $\mathcal{E}(\overline{F})$  with  $\overline{F}$  the complement of  $F$ . In some cases, typically for  $F$  sufficiently large, most matrices  $W \in \mathcal{E}(\overline{F})$  are uniquely determined by their eigenspaces. For those  $W \in \mathcal{E}(\overline{F})$ , there is no matrix  $A \in \mathcal{E}(F)$  non collinear with  $W$  that commutes with  $W$ . This property is encapsulated by the notion of *F-identifiability* as follows.

**Definition 1 ( $F$ -identifiability)** We say that a symmetric matrix  $W$  is  $F$ -identifiable if, and only if, the only solutions  $A$  with  $\text{Supp}(A) \subseteq \overline{F}$  to  $AW = WA$  are of the form  $A = \lambda W$  for some  $\lambda \in \mathbb{R}$ . Equivalently,

$$\{A \in \mathbb{R}^{N \times N} : A = A^\top, AW = WA \text{ and } \text{Supp}(A) \subseteq \overline{F}\} = \{\lambda W : \lambda \in \mathbb{R}\} \quad (1)$$

A matrix  $W$  is identifiable if the set of symmetric matrices with the same eigenvectors as  $W$  and whose support is included in  $\overline{F}$  is the line spanned by  $W$ .

**Remark 2** The dimension of the commutant, defined by

$$\text{Com}(W) := \{A \in \mathbb{R}^{N \times N} : A = A^\top, AW = WA\},$$

is entirely determined by the multiplicity of the eigenvalues of  $W$ . Indeed, letting  $\lambda_1, \dots, \lambda_s$  denote the different eigenvalues of  $W$  and  $\ell_1, \dots, \ell_s$  their multiplicities, one can show that

$$N \leq \dim(\text{Com}(W)) = \sum_{j=1}^s \frac{\ell_j(\ell_j + 1)}{2} \leq \frac{N(N+1)}{2}.$$

Now, the  $F$ -identifiability of  $W$  can be stated equivalently as  $\dim(\text{Com}(W) \cap \mathcal{E}(\overline{F})) = 1$ , observing that the left hand side of (1) is exactly  $\text{Com}(W) \cap \mathcal{E}(\overline{F})$ . Using a simple inclusion/exclusion formula, one can check that the condition

$$|F| \geq \dim(\text{Com}(W)) - 1$$

is necessary for the  $F$ -identifiability, where  $|F|$  denotes the cardinality of  $F$ . In particular, a matrix  $W$  with repeated eigenvalues requires a large set  $F$  of forbidden entries in order to be  $F$ -identifiable.

We have the following proposition.

**Proposition 3 (Lemma 2.1 in Barsotti et al. (2014))** Let  $S \subseteq \overline{F}$ , the set of  $F$ -identifiable matrices in  $\mathcal{E}(S)$  is either empty or a dense open subset of  $\mathcal{E}(S)$ .

This proposition conveys that the  $F$ -identifiability of a matrix  $W$  is essentially a condition on its support  $S$ . The proof uses the fact that non  $F$ -identifiable matrices in  $\mathcal{E}(S)$  can be expressed as the zeroes of a particular analytic function, we refer to Barsotti et al. (2014) for further details. By abuse of notation, we say that a support  $S \subseteq \overline{F}$  is  $F$ -identifiable if almost every matrix in  $\mathcal{E}(S)$  are  $F$ -identifiable.

Characterizing the  $F$ -identifiability appears to be a challenging issue since it can be viewed as understanding the eigen-structure of graphs through their common support. The particular case of the diagonal  $F_{\text{diag}}$  as the set of forbidden entries is given a particular attention in this paper. The  $F_{\text{diag}}$ -identifiability, or diagonal identifiability, can be reasonably assumed in many practical situations since it entails that  $W$  lives on a simple graph, with no self-loops. In Theorem 16 (see Appendix A.1), we introduce necessary and sufficient conditions on the target support  $\text{Supp}(W)$  for diagonal identifiability. Defining the kite graph  $\nabla_N$  of size  $N \geq 3$  as the graph  $(V, E)$  with vertices  $V = [1, N]$  and edges  $E = \{(k, k+1), 1 \leq k \leq N-1\} \cup \{(N-2, N)\}$  (see Figure 1), one simple sufficient condition on diagonal identifiability reads as follows, a proof is given in Section A.2.

**Proposition 4** If the graph  $G = ([1, N], S)$  contains the kite graph  $\nabla_N$  as a subgraph, then  $S$  is diagonally identifiable.

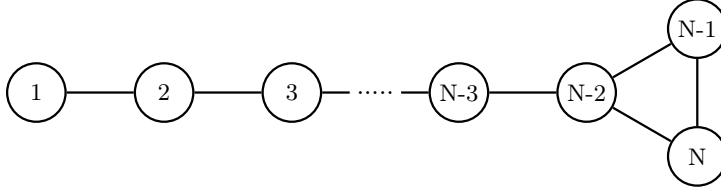


Figure 1: The kite graph  $\nabla_N$  on  $N$  vertices.

Denote  $G(N, p)$  the Erdős-Rényi model on graphs of size  $N$  where the edges are drawn independently with respect to the Bernoulli law of parameter  $p$ . Using Theorem 16, one can prove that  $\log N/N$  is a sharp threshold for diagonal identifiability in the Erdős-Rényi model (see Section A.4), it can be stated as follows.

**Theorem 5** *Diagonal identifiability in the Erdős-Rényi model occurs with a sharp phase transition with threshold function  $\log N/N$ : for any  $\varepsilon > 0$ , it holds*

- If  $p_N \geq (1 + \varepsilon) \log N/N$  and  $G_N \sim G(N, p_N)$  then the probability that  $\text{Supp}(G_N)$  is diagonally identifiable tends to 1 as  $N$  goes to infinity.
- If  $p_N \leq (1 - \varepsilon) \log N/N$  and  $G_N \sim G(N, p_N)$  then the probability that  $\text{Supp}(G_N)$  is diagonally identifiable tends to 0 as  $N$  goes to infinity.

In practice, one may expect that any target graph of size  $N$  with no self-loops and degree bounded from below by  $\log N$  is diagonally identifiable. In this case, it might be recovered from its eigenspaces. Conversely, small degree graphs (i.e. graphs with some vertices of degree much smaller than  $\log N$ ) may not be identifiable. In this case, there is no hope to reconstruct it from its eigenspaces since there exist another small degree undirected weighted graph with the same eigenspaces.

### 3. Some Concrete Models

#### 3.1 Markov chains

We begin with an example treated in the companion papers Barsotti et al. (2014, 2016). Consider a Markov chain  $(X_n)_{n \in \mathbb{N}}$  with finite state space  $[1, N]$  and transition matrix  $P \in \mathbb{R}^{N \times N}$ . Let  $(T_k)_{k \geq 1}$  be a sequence of random times such that  $T_{k+1} - T_k$  are i.i.d random variables independent of  $(X_n)_{n \in \mathbb{N}}$ . Denoting  $Y_k = X_{T_k}$ , remark that  $Y$  is also a Markov chain with transition matrix  $f(P)$  where  $f$  is the generating function of  $T_k - T_{k-1}$ . Therefore  $Q = f(P)$  may be estimated and one may recover  $P$  from  $Q$  without any information on the distribution of the time gaps.

#### 3.2 Vectorial AutoRegressive process

Consider a stationary Vectorial AutoRegressive process of order one  $(X_n)_{n \in \mathbb{Z}}$  verifying

$$X_{n+1} = W X_n + \varepsilon_n,$$

with  $\varepsilon_i$  i.i.d. random variables. Define as above  $Y_k = X_{T_k}$  where  $T_k$  are random times such that the time gaps  $T_k - T_{k-1}$  are i.i.d. with generating function  $f$ . Then, it holds

$$\mathbb{E}[Y_k | Y_{k-1}] = f(W) Y_{k-1},$$

which allows us to estimate  $K = f(W)$  and ultimately recover  $W$  from this estimate.

### 3.3 Ornstein-Uhlenbeck process

The same property holds for the continuous time version of this process, namely a vectorial Ornstein-Uhlenbeck process observed at random times verifying

$$dX_t = W X_t dt + dB_t.$$

In this case, one can check that, if  $Y_k = X_{T_k}$  where  $T_k$  are random times such that the time gaps  $T_k - T_{k-1}$  are i.i.d., then

$$\mathbb{E}[Y_k | Y_{k-1}] = f(e^{-W}) Y_{k-1},$$

where  $f$  is the characteristic function of the time gaps  $T_k - T_{k-1}$ .

### 3.4 Graphical models

Our model may be related to Graphical models, an overview can be found in [Verzelen \(2008\)](#) for instance. Indeed, one may consider  $W$  as the precision matrix, which is the inverse of the covariance matrix, having some non zero entries described by a graph of dependencies. Using  $f(x) = x^{-1}$ , this falls into our setting, trying to recover  $W$  from the estimation of the covariance matrix. Of course, in this case, it is better to use the knowledge of  $f$ , which certainly improves estimation. However, our procedure allows us to estimate the function  $f$  and heuristically validate the hypothesis  $f(x) = x^{-1}$ .

### 3.5 Seasonal VAR structure

We can also consider a toy example looking at a seasonal VAR structure without any randomness on the times of observations. Let  $T$  be a positive integer, and  $(u_k)_{k \in \mathbb{Z}}, (v_k)_{k \in \mathbb{Z}}$  be some periodic sequences of period  $T$ . Consider the following model

$$\forall k \in \mathbb{Z}, \quad Y_{k+1} = u_k Y_k + v_k W Y_k + \varepsilon_k.$$

We may observe the model only at time gap intervals  $T$  with some error, *i.e.*,  $X_t = Y_{tT+k_0} + \eta_t$ . This falls into the general frame

$$X_t = f(W) X_{t-1} + \mu_t,$$

where the  $\mu_t$  are time uncorrelated. In this case,  $K = f(W)$  can be estimated from the observations.

### 3.6 Spatial autoregressive gaussian fields

Note that gaussian autoregressive processes on  $\mathbb{Z}$  verify that the precision operator may be written as a polynomial of the adjacency operator of  $\mathbb{Z}$ . One natural way to extend this property (see for instance [Espinasse et al. \(2014\)](#)) is to define centered gaussian autoregressive fields on a graph through the same relation between the covariance operator  $K$  and the adjacency operator  $W$  (or the discrete Laplacian, depending on the framework) :  $K^{-1} = P(W)$ , with  $P$  a polynomial of degree  $d$ . In this framework, Graphical models methods will infer the graph of path of length  $d$ , whereas our methods aims to recover  $W$ . Note that this framework extends to ARMA spatial fields where  $K$  writes as a rational fraction of  $W$ , and the property of commutativity between  $W$  and  $K$  still holds.

In the previous cases, we assumed that we can not estimate directly  $W$ . For spatio-temporal processes, this means that we do not have access to a full trajectory. It may be the case when the sample is drawn at random times, or when we can only sample independently under stationary measure of such process, for instance when observation times are a lot larger than the typical evolution time's scale of the process. If the whole trajectory is available, it would be better to use this extra information, see for instance [Bento et al. \(2010\)](#) for the Ornstein-Uhlenbeck case and [Bento and Ibrahimi \(2014\)](#) for the non-linear diffusion case.

## 4. Estimating the Support

### 4.1 Empirical Contrast: the Commutator

The methodology presented in the paper relies on the fact that the target matrix  $W$  commutes with the matrix  $K$ . Indeed, recall that  $K := f(W)$ , see Section 2.1 for a definition of this notation. In particular, there exist an orthonormal matrix  $U$  and a diagonal matrix  $D$  with diagonal entries  $(\lambda_1, \dots, \lambda_1, \dots, \lambda_s, \dots, \lambda_s)$  with multiplicities  $(\ell_1, \dots, \ell_s)$  such that  $W = UDU^\top$  and  $K = Uf(D)U^\top$  where  $f(D)$  is a diagonal matrix with diagonal entries  $(f(\lambda_1), \dots, f(\lambda_1), \dots, f(\lambda_s), \dots, f(\lambda_s))$  and same multiplicities as above. Since  $f$  is assumed injective (and hence one to one on the spectrum of  $W$ ), the matrices  $W$  and  $K$  have exactly the same eigenspaces in the sense that the eigenspace  $E_{\lambda_k}(W)$  associated to  $\lambda_k$  (in the decomposition of  $W$ ) is exactly the one associated to  $f(\lambda_k)$  (in the decomposition of  $K$ ), namely

$$E_{f(\lambda_k)}(K) = E_{\lambda_k}(W) \quad (2)$$

and the dimension of this eigenspace is  $\ell_k$ , the multiplicity of  $\lambda_k$ . It follows that, when  $F$ -identifiability holds, the only solutions  $A$  with  $\text{Supp}(A) \subseteq \overline{F}$  to  $AK = KA$  are of the form  $A = tW$  for some  $t \in \mathbb{R}$ .

**Remark 6 (Reminder on matrix perturbation theory)** *Now, we do not observe  $K$  but a noisy version  $\hat{K}$ . For instance,  $\hat{K}$  is the estimation of  $K$  from a finite sample. The nice decomposition (2) does not hold anymore changing  $K$  by  $\hat{K}$ . But there exist an orthonormal matrix  $\hat{U}$ , a diagonal matrix  $\hat{D}$  with diagonal entries  $(\hat{\mu}_1, \dots, \hat{\mu}_N)$  such that  $\hat{K} = \hat{U}\hat{D}\hat{U}^\top$  and the following holds. Mirsky's inequality (Stewart and Sun, 1990, Corollary 4.12) and the Wedin's  $\sin(\theta)$  theorem (Stewart and Sun, 1990, P. 260) show that, for  $\hat{K}$  such that  $\|K - \hat{K}\|$  is sufficiently small (with respect to the minimal separation  $|f(\lambda_{k_1}) - f(\lambda_{k_2})|$  between distinct eigenvalues), then for all  $k = 1, \dots, s$ , the eigenvalues  $\hat{\mu}_{(\sum_{t=1}^{k-1} \ell_t)+1}, \dots, \hat{\mu}_{\sum_{t=1}^k \ell_t}$  are close to  $f(\lambda_k)$  and the space spanned by a group of eigenvectors, namely the vectors  $\hat{U}_{(\sum_{t=1}^{k-1} \ell_t)+1}, \dots, \hat{U}_{\sum_{t=1}^k \ell_t}$ , is close to  $E_{f(\lambda_k)}(K) = E_{\lambda_k}(W)$  (more precisely, the orthonormal projections onto these spaces are close in Frobenius norm).*

*If we consider  $A$  such that  $A\hat{K} = \hat{K}A$  then again these matrices share the same eigenspaces and we conclude that, up to label switching, the eigenvectors  $(V_k)_{k=1}^N$  of  $A$  are such that the spaces spanned by the group of eigenvectors  $V_{(\sum_{t=1}^{k-1} \ell_t)+1}, \dots, V_{\sum_{t=1}^k \ell_t}$  are close to the targets  $E_{\lambda_k}(W)$ , for  $k = 1, \dots, s$ .*

However, the choice  $A = W$  does not satisfy  $A\hat{K} = \hat{K}A$  and we need to relax this identity. Furthermore, remark that  $W\hat{K} - \hat{K}W = WE - EW$  denoting  $E = \hat{K} - K$  the estimation errors. It follows

$$\frac{\|W\hat{K} - \hat{K}W\|}{\|W\|} = \frac{\|WE - EW\|}{\|W\|} \leq 2\|E\|. \quad (3)$$

In view of (3) and of the discussion above, we use the following cost function

$$A \mapsto \frac{\|A\hat{K} - \hat{K}A\|}{\|A\|}, \quad A \in \mathcal{E}(\overline{F}) \setminus \{0\},$$

which aims at matrices for which the spaces spanned by some groups of its eigenspaces are close to the eigenspaces of the target. This empirical criterion was first used in Barsotti et al. (2014) in a similar context to reflect that  $W$  is expected to nearly commute with  $\hat{K}$ , provided that  $\hat{K}$  is sufficiently close to its true value  $K$ , see for instance (3).

## 4.2 The $\ell_0$ -approach

Given an estimator  $\widehat{K} = \widehat{K}_n$  of  $K$  build from a sample of size  $n$  and a set of forbidden entries  $F$  reflecting  $(\mathbf{H}_F)$ , we construct an estimator  $\widehat{S} = \widehat{S}_n$  of the target support  $S^* := \text{Supp}(W)$  as a minimizer of the criterion  $Q$  given by

$$\forall S \subseteq \overline{F}, \quad Q(S) := \min_{A \in \mathcal{E}(S) \setminus \{0\}} \frac{\|A\widehat{K} - \widehat{K}A\|}{\|A\|} + \lambda_n |S|,$$

for some tuning parameter  $\lambda_n > 0$  and defining the minimum of an empty set as  $\infty$ . Recall that  $\mathcal{E}(S)$  is the set of symmetric matrices  $A$  such that  $\text{Supp}(A) \subseteq S$ . Our estimator is

$$\widehat{S} \in \arg \min_{S \subseteq \overline{F}} Q(S)$$

Furthermore, we assume that the estimator  $\widehat{K}$  converges toward  $K$  in probability  $R_n$ , namely

$$\forall t > 0, \quad \mathbb{P}\{\|\widehat{K} - K\| \geq t\} \leq R_n(t), \quad (\mathbf{H}_2)$$

where  $t \mapsto R_n(t)$  is non-increasing and such that, for all  $t > 0$ ,  $R_n(t) \rightarrow 0$  as  $n$  goes to  $\infty$ .

**Theorem 7** *Assume that  $(\mathbf{H}_2)$  and  $(\mathbf{H}_F)$  hold. If  $W$  is  $F$ -identifiable, then*

$$\mathbb{P}\{\widehat{S} \neq S^*\} \leq R_n\left(\frac{c_0(S^*) - \lambda_n |S^*|}{4}\right) + R_n\left(\frac{\lambda_n}{2}\right),$$

where

$$c_0(S^*) := \min_{\substack{S \neq S^* \\ |S| \leq |S^*|}} \min_{A \in \mathcal{E}(S)} \frac{\|AK - KA\|}{\|A\|} > 0. \quad (4)$$

A proof of Theorem 7 is given in Section B.1.

**Corollary 8** *Under the assumptions of Theorem 7, if it holds*

$$\lambda_n \rightarrow 0 \quad \text{and} \quad \sum_{n \in \mathbb{N}} R_n\left(\frac{\lambda_n}{2}\right) < +\infty,$$

then one has  $\widehat{S} \rightarrow S^*$  almost surely.

Note that, based on the upper bound in Theorem 7, a good scaling may be  $\lambda_n^* = \frac{c_0(S^*)}{|S^*|+4}$  leading to the upper bound

$$\mathbb{P}\{\widehat{S} \neq S^*\} \leq 2R_n\left(\frac{c_0(S^*)}{2|S^*|+8}\right) \xrightarrow{n \rightarrow \infty} 0$$

which is optimal up to a constant less than 2. Interestingly, this oracle choice  $\lambda_n^*$  does not depend on  $n$  but this calibration is not relevant since both  $c_0(S^*)$  and  $|S^*|$  are unknown. Alternatively, we may choose a sequence  $\lambda_n$  decreasing slowly to 0 to ensure both conditions of Corollary 8.

## 4.3 Edge significance based on the commutator criterion

The  $\ell_0$ -approach meets with the curse of dimensionality. In practice, a backward methodology provides a computationally feasible alternative to the support reconstruction problem. Starting from the maximal acceptable support  $\overline{F}$ , the idea of the backward procedure is to remove the least significant entries one at a time and stop when every entry is significant. Using the corresponding small case letter to denote the vectorization of a matrix, *e.g.*,  $a = \text{vec}(A) =$



$(A_{11}, \dots, A_{N1}, \dots, A_{1N}, \dots, A_{NN})^\top$ , significancy can be leveraged using the Frobenius norm of the commutator operator  $a \mapsto \Delta(K)a = \text{vec}(KA - AK)$ , where

$$\Delta(K) = I \otimes K - K \otimes I \in \mathbb{R}^{N^2 \times N^2}$$

and  $\otimes$  denotes the Kronecker product. Indeed, searching for the target  $W$  in the commutant of  $K$  reduces to searching for  $w = \text{vec}(W)$  in  $\ker(\Delta(K))$ , the kernel of  $\Delta(K)$ . Because the Frobenius norm coincides with the Euclidean norm of the vectorization, the functions  $A \mapsto \|\hat{K}A - A\hat{K}\|^2$  and  $a \mapsto \|\Delta(\hat{K})a\|^2$  can be used indistinctly as cost functions. Minimizing this criterion over model spaces of decreasing size, we may consider sequences of least-squares estimates in the sequel.

#### ASSUMPTIONS

Assume the three following hypotheses  $(\mathbf{H}_\Sigma)$ ,  $(\mathbf{H}_1)$  and  $(\mathbf{H}_{\text{Id}})$ .

◦ Deriving the asymptotic law of least-squares estimators, we may assume that the estimate  $\hat{K}$  is such that

$$\sqrt{n}(\hat{k} - k) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \Sigma), \quad (\mathbf{H}_\Sigma)$$

where  $\Sigma$  is a  $N^2 \times N^2$  covariance matrix (either known or that can be estimated). For instance, one can think of  $\hat{K}$  as the empirical covariance when observing a sample of vectors of covariance  $K$ . This condition is verified for instance in the framework considered in Barsotti et al. (2014, 2016). Note that asymptotic normality is a standard ground base investigating any least-squares procedure.

◦ In order to exclude the trivial solution  $a = 0$ , the target  $W$  is assumed normalized

$$\mathbf{1}^\top w = 1, \quad (\mathbf{H}_1)$$

where  $\mathbf{1}$  has all its entries equal to one. Because the available information on  $W$  is of spectral nature and as such, is scale-invariant, a normalization of some kind is crucial for the identifiability. Here, the condition  $\mathbf{1}^\top w = 1$  achieves two goals: preventing the null matrix form being a solution and making the problem identifiable.

**Remark 9** *The main drawback of this normalization concerns the situation where the entries of  $W$  sum up to zero, in which case the normalization is impossible. If the context suggests that the solution may be such that  $\mathbf{1}^\top w = 0$ , a different affine normalization  $\mathbf{v}^\top w = 1$  (with any fixed vector  $\mathbf{v}$ ) must be used, without major changes in the methodology. In practice, one may consider the vector  $\mathbf{v}$  at random (for instance with isotropic law), so that  $(\mathbf{H}_1)$  is almost surely fulfilled for any fixed target  $w$ .*

Observe that if one knows in advance that the target  $W$  has nonnegative entries then the normalization  $(\mathbf{H}_1)$  is acceptable.

◦ For  $S$  a support included in  $\bar{F}$ , we aim at a solution in the affine space

$$\mathcal{A}_S := \{a = \text{vec}(A) : \text{Supp}(A) \subseteq S, A = A^\top, \mathbf{1}^\top a = 1\}.$$

with linear difference space given by

$$\mathcal{L}_S := \{a = \text{vec}(A) : \text{Supp}(A) \subseteq S, A = A^\top, \mathbf{1}^\top a = 0\}.$$

By abuse of notation,  $\mathcal{A}_S$  may refer both to the space of matrices or their vectorizations. To find the target support  $S^*$ , one must exploit the fact that the vector  $w$  lies in the intersection of  $\ker(\Delta(K))$  and  $\mathcal{A}_{\bar{F}}$ . Actually,  $w$  can then be recovered if the intersection is reduced to the singleton  $\{w\}$ . In this case, the matrix  $W$  and its support  $S^*$  are  $F$ -identifiable. Hence, we assume that

$$\ker(\Delta(K)) \cap \mathcal{L}_{\bar{F}} = \{0\}, \quad (\mathbf{H}_{\text{Id}})$$

which is implied by  $F$ -identifiability, see Definition 1.



## ASYMPTOTIC NORMALITY AND A SIGNIFICANCE TEST

The framework under consideration can be viewed as a heteroscedastic linear regression model with noisy design for which  $w = \text{vec}(\mathbf{W})$  is the parameter of interest. Indeed, consider for each support  $S \subseteq \overline{F}$  a full-ranked matrix  $\Phi_S \in \mathbb{R}^{N^2 \times \dim(\mathcal{A}_S)}$  whose column vectors form a basis of  $\mathcal{L}_S$ . Assuming that  $\mathbf{W}$  is  $F$ -identifiable and taking  $S \subseteq \overline{F}$ , the operator  $\Delta(\mathbf{K})\Phi_S$  is one-to-one. In this case, evaluating the commutator  $a \mapsto \Delta(\mathbf{K})a$  over  $\mathcal{A}_S$  reduces to considering the map

$$b \mapsto \Delta(\mathbf{K})(a_0 - \Phi_S b), \quad b \in \mathbb{R}^{\dim(\mathcal{A}_S)},$$

with  $a_0$  chosen arbitrarily in  $\mathcal{A}_S$ . When replacing the unknown  $\Delta(\mathbf{K})$  with its estimate  $\Delta(\widehat{\mathbf{K}})$ , the minimization of the criterion  $a \mapsto \|\Delta(\widehat{\mathbf{K}})a\|^2$  over  $\mathcal{A}_S$  can be written similarly as a linear regression framework where the parameter of interest is estimated by

$$\widehat{\beta}_S \in \arg \min_{b \in \mathbb{R}^{\dim(\mathcal{A}_S)}} \|\Delta(\widehat{\mathbf{K}})(a_0 - \Phi_S b)\|^2. \quad (5)$$

We recognize a linear model with response  $y = \Delta(\widehat{\mathbf{K}})a_0$  and noisy design matrix  $X = \Delta(\widehat{\mathbf{K}})\Phi_S$ . In this setting, remark that  $w = a_0 - \Phi_S \beta$  with  $\beta$  the unique solution to  $\Delta(\mathbf{K})(a_0 - \Phi_S \beta) = 0$ . Denoting by  $\mathbf{M}^\dagger$  the pseudo-inverse of a matrix  $\mathbf{M}$ , we deduce the following result.

**Theorem 10** *If  $S^* \subseteq S$ , the estimator  $\widehat{\beta}_S$  is asymptotically Gaussian with*

$$\sqrt{n}(\widehat{\beta}_S - \beta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \Omega_S),$$

where  $\Omega_S := (\Phi_S^\top \Delta(\mathbf{K}))^\dagger \Delta(\mathbf{W}) \Sigma \Delta(\mathbf{W}) (\Delta(\mathbf{K}) \Phi_S)^\dagger$ .

We then have

$$\widehat{w}_S = \text{vec}(\widehat{\mathbf{W}}_S) = \arg \min_{a \in \mathcal{A}_S} \|\Delta(\widehat{\mathbf{K}})a\|^2 = a_0 - \Phi_S \widehat{\beta}_S. \quad (6)$$

The asymptotic distribution of  $\widehat{w}_S$  follows directly from Theorem 10,

$$\sqrt{n}(\widehat{w}_S - w) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \Phi_S \Omega_S \Phi_S^\top). \quad (7)$$

The limit covariance matrix is unknown, but plugging the estimates  $\widehat{\mathbf{W}}_S$ ,  $\widehat{\mathbf{K}}$  and  $\widehat{\Sigma}$  yields an estimator  $\Phi_S \widehat{\Omega}_S \Phi_S^\top$ , which is consistent under the  $F$ -identifiability assumption. In particular, the diagonal entry of  $\Phi_S \widehat{\Omega}_S \Phi_S^\top$  associated to the  $(i, j)$ -entry of  $\mathbf{W}$ , which we denote  $\widehat{\sigma}_{S,ij}^2$ , provides a consistent estimator for the asymptotic variance of  $\widehat{\mathbf{W}}_{S,ij}$ . As a result, the statistic

$$\tau_{ij}(S) := \sqrt{n} \frac{\widehat{\mathbf{W}}_{S,ij}}{\widehat{\sigma}_{S,ij}} \quad (8)$$

can be used to measure the relative significance of the estimated entry  $\widehat{\mathbf{W}}_{S,ij}$ . The backward support selection procedure is then implemented by the recursive algorithm as follows.

---

### Algorithm 1: Backward algorithm for support selection

---

**Data:** A set of forbidden entries  $F$ , a matrix  $\widehat{\mathbf{K}}$ .

**Result:** A sequence of estimators  $\widehat{\mathbf{W}}_{S_1}, \widehat{\mathbf{W}}_{S_2}, \dots$  with nested supports  $S_1 \supset S_2 \supset \dots$

- 1: Start with the maximal acceptable support  $S_1 = \overline{F}$ ,
  - 2: At each step  $k$ , compute the statistics  $\tau_{ij}(S_k)$  for all  $(i, j) \in S_k$ ,
  - 3: Remove the least significant edge  $(i, j)$  which minimizes  $|\tau_{ij}(S_k)|$  for  $(i, j) \in S_k$ , and set  $S_{k+1} = S_k \setminus \{(i, j), (j, i)\}$ ,
  - 4: Stop when all edges have been removed.
-

The backward algorithm produces a sequence of nested supports that one can choose to stop once all the edges are judged significant, that is, when all the statistics  $\tau_{ij}(S_k)$ ,  $(i, j) \in S_k$  exceed in absolute value some fixed threshold  $\tau_0$ . Owing to the asymptotic normality of  $\widehat{W}_{S,ij}$  shown in Eq. (7), the  $(1 - \frac{\alpha}{2})$ -quantile of the standard Gaussian distribution would appear as a reasonable choice for the threshold  $\tau_0$ , as it boils down to performing an asymptotic significance test of level  $\alpha$ . However, due to the slow convergence to the limit distribution and the tendency to overestimate the variance for small sample sizes (see Figure 2), a threshold based on the Gaussian quantile inevitably leads to an overly large estimated support. Nevertheless, we show that an adaptive calibration of the threshold can be achieved by considering the overall behavior of the commutator  $\Delta(\widehat{K})\widehat{w}_{S_m}$  computed over the nested sequence of active supports.

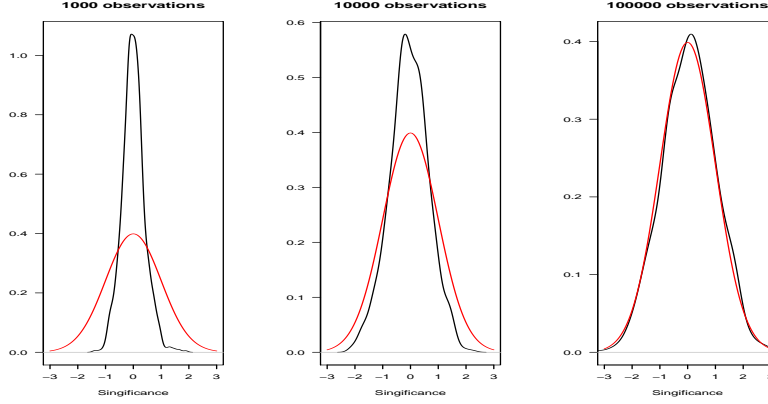


Figure 2: Estimated density of the statistic  $\tau_{ij}(S)$  for an edge  $(i, j) \in S \setminus S^*$  compared to its theoretical Gaussian limit distribution, for samples of size  $n = 1000$  (left),  $n = 10000$  (center) and  $n = 100000$  (right).

#### CALIBRATION OF THE THRESHOLD BY CROSS-VALIDATION

By removing the least significant edge at each step, the backward algorithm generates a sequence of nested active supports  $S_1 \supset \dots \supset S_\ell$ , that we refer to as a “trajectory”. Along this trajectory, we compute the empirical contrast defined by

$$\forall S \subseteq \overline{F}, \quad S \mapsto \text{Crit}(\widehat{W}_S, \widehat{K}) := \frac{\|\widehat{W}_S \widehat{K} - \widehat{K} \widehat{W}_S\|}{\|\widehat{W}_S\|}. \quad (9)$$

Note that computing this criterion boils down to compute  $\widehat{W}_S$  which is a simple projection onto  $\mathcal{A}_S$  as shown in (6).

When the true support  $S^*$  lies in the trajectory, one expects to observe a “gap” in the sequence  $j \mapsto \text{Crit}(\widehat{W}_{S_j}, \widehat{K})$  when  $S_j$  goes from  $S^*$  to a smaller support. Indeed:

- For  $S^* \subseteq S$ , the target  $W$  is consistently estimated by  $\widehat{W}_S$  so that  $\text{Crit}(\widehat{W}_S, \widehat{K})$  tends to zero at rate  $\sqrt{n}$ ,
- For  $S \subsetneq S^*$ , the lower bound  $\|A\widehat{K} - \widehat{K}A\| \geq \|AK - KA\| - 2\|\widehat{K} - K\|\|A\|$  yields

$$\text{Crit}(\widehat{W}_S, \widehat{K}) = \frac{\|\widehat{W}_S \widehat{K} - \widehat{K} \widehat{W}_S\|}{\|\widehat{W}_S\|} \geq c(S) - 2\|\widehat{K} - K\| \quad (10)$$

with  $c(S) := \min_{A \in \mathcal{A}_S} \|\mathbf{A}\mathbf{K} - \mathbf{K}\mathbf{A}\| / \|\mathbf{A}\|$  a positive constant. In particular, one has

$$\min_{S \subsetneq S^*} c(S) \geq \min_{\substack{S \neq S^* \\ |S| \leq |S^*|}} c(S) = c_0(S^*) > 0$$

where  $c_0(S^*)$  is defined in (4).

In some way,  $c_0(S^*)$  measures the amplitude of the signal: one expects to be able to recover the target  $\mathbf{W}$  when the estimation error  $\|\hat{\mathbf{K}} - \mathbf{K}\|$  reaches at least the same order as  $c_0(S^*)$ . The true support  $S^*$  then corresponds to a transitional gap in the contrast curve that can be captured by a suitably chosen threshold  $t > 0$ . Since  $\hat{\mathbf{K}}$  converges toward  $\mathbf{K}$  in probability, any threshold  $0 < t < c_0(S^*)$  will work with probability one asymptotically.

**Remark 11** *The condition that  $S^*$  lies in the trajectory of nested supports is crucial to detect the commutation gap, although seldom verified in practice due to the tremendous amount of testable supports. This issue is specifically targeted by the boosted version of the backward algorithm discussed in Section 4.4.*

An obstacle to the detection of the commutation gap is the increasing behavior of the commutator over the nested trajectory  $S_1 \supset \dots \supset S_\ell$ . This phenomenon, indirectly caused by the dependence between the trajectory and  $\hat{\mathbf{K}}$ , can be annihilated when considering the empirical contrast over a trajectory built from a training sample. In fact, the monotonicity can even be “reversed” before reaching the true support if the  $\tilde{\mathbf{W}}_{S_j}$  are estimated independently from  $\hat{\mathbf{K}}$ . This can be explained as follows: Consider the ideal scenario where estimators  $\tilde{\mathbf{W}}_{S_1}, \dots, \tilde{\mathbf{W}}_{S_\ell}$  are built from the backward algorithm using an estimator  $\tilde{\mathbf{K}}$  independent from  $\hat{\mathbf{K}}$ . We assume moreover that the true support  $S^*$  lies in the trajectory  $S_1 \supset \dots \supset S_\ell$ . The trick is to write

$$\Delta(\hat{\mathbf{K}})\tilde{w}_{S_j} = \Delta(\tilde{\mathbf{K}})w + \Delta(\mathbf{K})\tilde{w}_{S_j} + \Delta(\hat{\mathbf{K}} - \mathbf{K})(\tilde{w}_{S_j} - w),$$

and to analyze the three terms separately:

- The term  $\Delta(\hat{\mathbf{K}})w$  has no influence as it is common to all supports in the trajectory.
- The term  $\Delta(\mathbf{K})\tilde{w}_{S_j}$  approaches zero as  $\tilde{w}_{S_j}$  gets closer to  $w$ . Heuristically, the variance of  $\tilde{w}_{S_j}$ , and incidentally that of  $\Delta(\mathbf{K})\tilde{w}_{S_j}$ , is larger for over-fitting supports  $S \supsetneq S^*$ . This results in the sequence  $j \mapsto \Delta(\mathbf{K})\tilde{w}_{S_j}$  being stochastically decreasing as  $S_j$  approaches  $S^*$  from above. On the other hand, the bias is expected to dominate once the trajectory passes through the true value  $S^*$ , making the remaining of the sequence  $\Delta(\mathbf{K})\tilde{w}_{S_j}$  increase stochastically.
- The term  $\Delta(\hat{\mathbf{K}} - \mathbf{K})(\tilde{w}_{S_j} - w)$  is negligible for  $S \supsetneq S^*$ , as both  $\hat{\mathbf{K}} - \mathbf{K}$  and  $\tilde{w}_{S_j} - w$  tend to zero independently. We emphasize that this argument no longer holds without the independence of  $\tilde{w}_{S_j}$  and  $\hat{\mathbf{K}}$ . This is precisely why we use a training sample.

Thus, the sequence  $j \mapsto \text{Crit}(\tilde{\mathbf{W}}_{S_j}, \hat{\mathbf{K}}) = \|\Delta(\hat{\mathbf{K}})\tilde{w}_{S_j}\| / \|\tilde{w}_{S_j}\|$  is expected to achieve its minimum for the best estimator  $\tilde{w}_{S_j}$  in the trajectory, that is for  $S_j = S^*$ . Furthermore, beyond the true support (for small active supports),  $\tilde{w}_{S_j}$  is not a consistent estimator of  $w$  so that the criterion no longer approaches zero, resulting in the so-called commutation gap.

The “reversed” monotonicity provides an easy way to calibrate the threshold in the backward algorithm. Indeed, since  $S_j \mapsto \Delta(\hat{\mathbf{K}})\tilde{w}_{S_j}$  is expected to decrease when approaching the true support (coming from larger active supports along a trajectory), the estimated support can be heuristically chosen as the last time the criterion is below an adaptive threshold, see Figure 3. In particular,  $\text{Crit}(\tilde{\mathbf{W}}_{S_1}, \hat{\mathbf{K}})$  can be used as an adaptive threshold for the backward algorithm when the estimator  $\hat{\mathbf{K}}$  and the trajectory  $S_1 \supset \dots \supset S_\ell$  are obtained from independent samples.

Of course, to afford splitting the sample to build the  $\tilde{W}_{S_j}$  independent from  $\hat{K}$  may be unrealistic. Nevertheless, the numerical study suggests that the independence is well mimicked when  $\hat{K}$  is built from the whole dataset but the backward algorithm sequence  $\tilde{W}_{S_1}, \dots, \tilde{W}_{S_\ell}$  is obtained from a learning sub-sample, as illustrated in Figure 3. Empirically, the optimal size of training samples could be calibrated in function of the number of observations using the robustness of the outputs of the algorithm. In this paper, we always draw training samples by taking each observation with probability 1/2, with no consideration regarding the size of the whole sample.

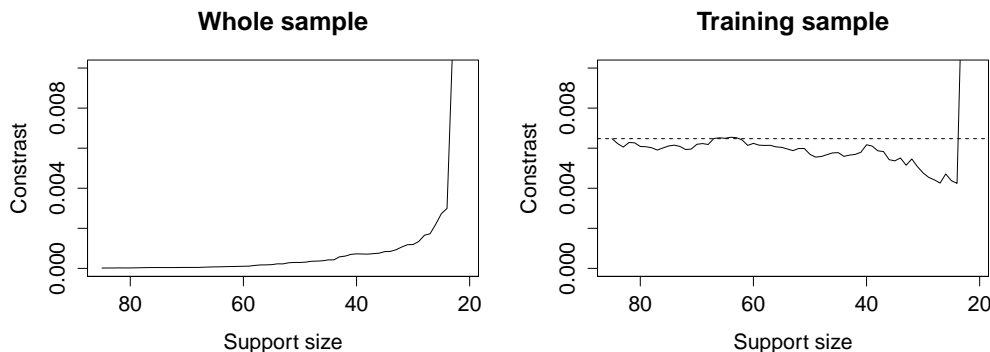


Figure 3: The contrast sequence  $j \mapsto \text{Crit}(\tilde{W}_{S_j}, \hat{K})$  computed in the example of Section 5.2. The nested support sequence and estimators  $\tilde{W}_{S_j}$  are obtained from the backward algorithm implemented on the whole sample (left) and on a training sample of half size (right). In both cases,  $\hat{K}$  is constructed from the whole sample. Using a training sample manages to reverse the monotonicity in the first part of the sequence, thus making the commutation gap easier to locate. The initial value of the sequence  $t = \text{Crit}(\tilde{W}_{S_1}, \hat{K})$  then provides a tractable adaptive choice for the threshold.

#### 4.4 Boosting of the backward algorithm

The main weakness of the backward procedure remains that it requires the true support  $S^*$  to lie in the trajectory  $S_1 \supset \dots \supset S_\ell$  obtained from removing the least significant edge one at a time. In practice, this condition is rarely verified, especially with small datasets. A way to solve this issue is to replicate the backward algorithm over a collection of random sub-samples, a process commonly known to as *boosting*. The description of this algorithm is given in Algorithm 2.

The boosted algorithm produces a collection of estimated supports in a way to make the final decision more robust. At this point, several solutions are possible: select the most represented support among the  $\hat{S}_m$ 's, keep the edges present in the most supports etc... A preliminary detection of the outliers among the  $\hat{S}_m$ 's, e.g. by removing beforehand the supports  $\hat{S}_m$ 's that are either too big or too small, might also considerably improve the method, as we illustrate on actual examples in Section 5.

### 5. Numerical study

#### 5.1 Toy example

In the previous section, we have introduced different algorithms. To emphasize the motivation of the boosting algorithm, we consider a simple example, and implement the different algorithms

---

**Algorithm 2:** Boosted backward algorithm

---

**Data:** A set of forbidden entries  $F$ , a sample  $X$ .

**Result:** A collection of estimated supports  $\hat{S}_m, m = 1, \dots, M$ .

- 1: Build  $M$  bootstrapped samples without replacement.
- 2: For each sub-sample  $m = 1, \dots, M$ , build an estimator  $\tilde{K}_m$  of  $K$ .
- 3: For all  $m$ , run Algorithm 1 without stopping condition and return  $M$  trajectories  $S_{1m} \supset \dots \supset S_{\ell m}$  and the corresponding estimators  $\tilde{W}_{S_{km}}$ .
- 4: Evaluate the empirical contrast  $\text{Crit}(\tilde{W}_{S_{km}}, \hat{K})$  over each trajectory with the estimator  $\hat{K}$  calculated from the whole sample.
- 5: For each trajectory, return the estimated support  $\hat{S}_m := S_{\hat{k}_m m}$  as the last support whose contrast lies below the initial value:

$$\hat{k}_m := \max \{k = 1, \dots, \ell : \text{Crit}(\tilde{W}_{S_{km}}, \hat{K}) \leq \text{Crit}(\tilde{W}_{S_{1m}}, \hat{K})\}.$$

---

for support recovery. To check the performances of the  $\ell_0$  procedure, we need to consider a graph with a small number of vertices (since the  $\ell_0$  complexity grows with  $2^{N(N-1)/2}$  where  $N$  denotes the number of vertices). Here, we consider the graph  $G_1$  represented in Figure 4, the kite graph on 5 vertices.

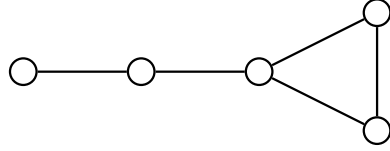


Figure 4: The kite graph  $G_1 = \nabla_5$ .

We choose  $W$  as the (normalized) adjacency matrix of  $G_1$  then draw a sample of size  $n = 500$  of centered Gaussian vectors  $X_1, \dots, X_n$  of  $\mathbb{R}^5$  with covariance matrix  $K = \exp(W)$ . We assume known that  $G_1$  contains no self-loop so that we take  $F = F_{diag}$  as the set of forbidden values. In this simple example, the constant  $c_0(S^*)$  (see Eq. (4)) can be calculated explicitly, yielding  $c_0(S^*) \approx 0.12$ . In comparison, for  $n = 500$ ,  $\mathbb{E}\|\hat{K} - K\|$  is evaluated to approximately 0.27 by Monte-Carlo. We expect to be able to recover the true support when the noise level drops below the signal amplitude. Based on the bound of Eq. (10), this occurs as soon as  $\|\hat{K} - K\| \leq c_0(S^*)/2$  however, because this bound is not sharp, a lesser level of precision is required in practice.

We compare the following algorithms:

1. *Contrast penalized  $\ell_0$  minimization with optimal penalization constant.* We compute

$$\hat{S} = \arg \min_{S \subseteq \bar{F}_{diag}} \left\{ \min_{A \in \mathcal{E}(S) \setminus \{0\}} \frac{\|A\hat{K} - \hat{K}A\|}{\|A\|} + \lambda|S| \right\}.$$

The constant  $\lambda$  is chosen as the best possible value, minimizing the oracle error  $\delta(\hat{S})$  measured by the symmetric difference between  $\hat{S}$  and  $S^*$ :  $\delta(\hat{S}) = |\hat{S} \cup S^* \setminus \hat{S} \cap S^*|$ . Because the calibration parameter  $\lambda$  is chosen optimally for each realization of  $\hat{K}$ , the numerical performances of the method can be expected to be overestimated compared to a fully data-driven procedure.

2. *Thresholding contrast minimization with optimal threshold.* The target matrix  $W$  is estimated over the maximal acceptable support  $\bar{F}_{diag}$ . We then compute

$$\hat{S} = \{(i, j) : |\hat{W}_{ij}| > t\},$$

where the threshold  $t$  is chosen so as to minimize the oracle error  $\delta(\hat{S})$  for each realization of  $\hat{K}$ .

3. *Backward algorithm.* We generate a training sample by taking each observation with probability  $1/2$  independently, from the whole sample. The estimator of  $K$  in this sub-sample is denoted  $\tilde{K}$ . We implement Algorithm 1 on  $\tilde{K}$ , yielding a trajectory  $S_1 \supset \dots \supset S_\ell$  of nested supports whose sizes vary from  $|S_1| = 20$  (the full off-diagonal support) to  $|S_\ell| = 12$  (the minimal size required for diagonal identifiability), along with the associated estimators  $\tilde{W}_{S_k}$ ,  $k = 1, \dots, \ell$ . Remark that because the supports are symmetric, two entries are removed at each step so that  $\ell = 5$  in this case. We then compute the threshold  $t = \text{Crit}(\tilde{W}_{S_1}, \tilde{K})$  corresponding to the initial value of the contrast. The estimated support  $\hat{S}$  is defined as the smallest support  $S$  in the trajectory such that  $\text{Crit}(\tilde{W}_S, \tilde{K}) \leq t$ .
4. *Boosted backward algorithm.* The previous algorithm is implemented over  $M = 100$  training samples drawn keeping observations with probability  $1/2$ . For each  $m = 1, \dots, M$ , we retain
- the threshold  $t_m = \text{Crit}(\hat{W}_{S_{1m}}, \hat{K})$  corresponding to the initial value of the contrast,
  - the estimated support, that is, the smallest support  $\hat{S}_m$  in the trajectory such that  $\text{Crit}(\hat{W}_{\hat{S}_m}, \hat{K}) \leq t_m$ .

The final decision  $\hat{S}$  is obtained as follows. Only a proportion  $q$  of the training samples  $m$  with a small initial contrast  $t_m$ , which are expected to provide more accurate results, are kept (in the whole paper, we chose  $q = 2/\sqrt{M}$  empirically). Then, the smallest support among the remaining candidates is retained, choosing one at random if it is not unique.

**Remark 12** *We can view our problem as a linear regression such that the observation that we aim at regressing is null,  $y = 0$  and the design operator  $A \mapsto \hat{K}A - A\hat{K}$  is noisy. Our goal is to find a solution in the kernel of the operator  $A \mapsto KA - AK$ . In this context, a Lasso procedure (i.e., minimizing  $\|\hat{K}A - A\hat{K}\|^2 + \lambda\|A\|_1$ ) without further constraints leads to the null matrix solution. Therefore, we need to add a condition to avoid the null solution  $\hat{W} = 0$ . Since we can only recover the target up to a scaling parameter, we should consider for instance that  $\|W\| = 1$  and adding the constraint  $\|A\| = 1$ . It results in a non-convex program with no guarantees that a local minimum is the solution to the program.*

*Recall that we aim at recovering the exact support when the number of observations is large. But using the  $\ell_1$  penalty tends to overestimate the support and any conservative choice of  $\lambda$  will lead to false positives in the estimated support. Furthermore, it can be understood that a full matrix may commute with  $\hat{K}$ , and at the same time it may have a small  $\ell_1$  norm. That is to say that there may be no restricted eigenvalue condition for the noisy design operator in our framework.*

*Hence, when aiming for support recovery, the typical solution is to vanish the small entries of  $\hat{W}$ , making it no more efficient than the thresholded  $\ell_2$  procedure considered in Algorithm 2. For this reason, the numerical performances of the Lasso procedure are not included in the study.*

The next table compares the performances of the four algorithms. We calculated the Monte-Carlo estimated mean error  $\mathbb{E}(\delta(\hat{S}))$  and probability of exact recovery  $\mathbb{P}\{\hat{S} = S^*\}$  for 1000 repetitions of the experiment. The average computational time (obtained with the function

timer of Scilab) on a processor Intel Xeon @2.6 GHz are shown, using the oracle values of  $\lambda$  and  $t$  for the first two algorithms (the calibration of these parameters is thus not accounted for in the computation time).

Algorithm	$\ell_0$	$\ell_2$ -thresholding	Backward	Boosted Backward
Mean Error	0.45	0.37	1.95	0.68
Exact recovery	68%	75%	23%	61%
CPU time (s)	0.32	0.002	0.009	0.59

In this example, the first two algorithms are the more accurate. The percentage of successful recoveries for the boosted backward algorithm is nonetheless competitive given that the first two procedures have been calibrated optimally for each experiment, which would be highly infeasible in practice. Finally, we observe that although it is much more expensive computationally, the boosted version of the backward algorithm yields an undeniable improvement.

Upper bounds for the time and space complexity of the algorithms are given in the next table. The time complexity is calculated as the number of different supports  $S$  considered to lead to the solution in function of the size  $N$  of the graph and the number  $M$  of training samples. The spatial complexity measures the memory size needed to compute the solution. In this setting, it is the main limitation for applying the procedures to large graphs. The  $N^4$  comes from the computation of  $\Delta(K) = K \otimes I - I \otimes K$  in the solver. Admittedly, the complexity could be improved by using sparse matrix encoding although this was not implemented.

Algorithm	$\ell_0$	$\ell_2$ -thresholding	Backward	Boosted Backward
Space Complexity	$O(N^4)$	$O(N^4)$	$O(N^4)$	$O(N^4)$
Time Complexity	$O(2^{N(N-1)/2})$	$O(1)$	$O(N^2)$	$O(N^2.M)$

On the current version, the boosted backward algorithm contains scalability issues for big graphs due to its space complexity. Leads to reduce the spatial complexity include using sparse matrix encoding or the use of cheap approximations of the criterion. These shall be investigated in future works.

## 5.2 A diagonally identifiable matrix

The advantages of the boosted backward algorithm are highlighted for larger graphs. In the next example, we consider the graph  $G_2$  on  $N = 15$  vertices represented in Figure 5. The experimental conditions are similar to that of the previous example, a sample of size  $n = 10000$  is drawn from a centered Gaussian vector of variance  $K = \exp(W)$  where  $W$  is the normalized adjacency matrix of  $G_2$ , with normalizing constant chosen such that  $\mathbf{1}^\top w = 1$ . The implementation of the different algorithms follow the description of the previous example.

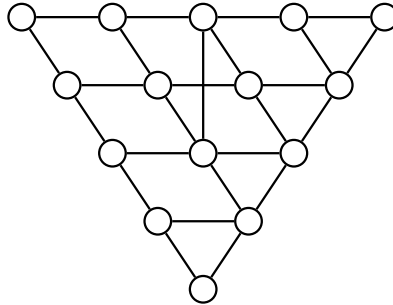


Figure 5: The graph  $G_2$  is diagonally identifiable.



In this case, the number of possible supports is too large for the  $\ell_0$  method to be implementable while the accuracy of the thresholded  $\ell_2$  drops considerably compared to smaller cases. We summarize the results in the following table.

Algorithm	$\ell_2$ -thresholding	Backward	Boosted Backward
Mean Error	10	25	1
Exact recovery	22%	26%	69%
CPU time (s)	0.04	2.5	256

A drawback of the boosted backward algorithm is the larger computational time: it takes around 4 minutes in average to estimate the support. Being essentially  $M = 100$  repetitions of the backward algorithm, the numerical complexity of the boosted version is roughly  $M$  times that of the simple backward algorithm, although the improvement is, here again, clear.

To illustrate the influence of the unknown function  $f$ , we consider  $f : t \mapsto (1 - t)^{-2}$  and reproduce the numerical study for  $K = f(W)$ . The results for various sample sizes are gathered in the next table, for  $M = 100$  boosting runs.

$n$	10000	5000	2000	1000
Exact recovery	97%	87%	83%	13%
Mean error	0.05	0.33	0.9	8.5

The probability of recovering the true support appears to be greater than in the previous example (97% against 69% previously for  $n = 10000$ ). This sheds lights on another important factor in the efficiency of the methods which is the separability of the spectrum of  $K$ . Indeed, in this framework, the information needed to recover  $W$  lies in its eigenspaces, which are estimated via  $\hat{K}$ . The accuracy of these estimates depends on the distance between the different eigenvalues (see e.g. Corollary 4.12 in [Stewart and Sun \(1990\)](#) and the Wedin'  $\sin(\theta)$  theorem in [Stewart and Sun \(1990\)](#)). Thus, for  $\lambda_1, \dots, \lambda_N$  the spectrum of  $W$ , the ability to recover  $W$  from  $K = f(W)$  is strongly impacted by the distances  $|f(\lambda_i) - f(\lambda_j)|, i, j = 1, \dots, N$ . In this situation where  $f$  is the exponential function,  $W$  having few negative eigenvalues will thus have a positive impact on the estimation. For the sake of comparison, the spectrum of  $W$ , given by  $\{-0.45, -0.28, -0.26, -0.21, -0.18, -0.16, -0.08, -0.01, 0.03, 0.10, 0.14, 0.19, 0.31, 0.34, 0.52\}$ , is more "spread" by the function  $t \mapsto (1 - t)^{-2}$  than by the exponential, see Figure 6.

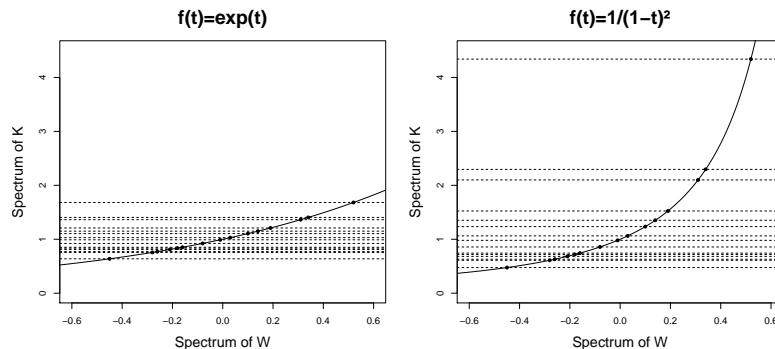


Figure 6: Separability of the spectrum of  $K = f(W)$  for  $f : t \mapsto \exp(t)$  (left) and  $f : t \mapsto (1 - t)^{-2}$  (right). The eigenvalues of  $K$  are more separated in the second case, making it easier to approximate its eigenspaces from the estimator  $\hat{K}$ .

**Remark 13** We also implemented the procedure in a random setting where  $W$  is drawn from an Erdős-Rényi graph with binomial entries. The conclusions obtained in this case are similar to those already discussed and shall not be presented to avoid redundancy.

## 6. Real life application

We now implement the boosted backward algorithm on real life data provided by Météorage and Météo France. The data contain the daily number of lightnings during a 3 year period in 16 regions of France localized on a  $4 \times 4$  grid. We expect to recover the spatial structure of the graph from the dependence of the lightning occurrences between the regions.

The data are refined as follows. We first eliminate day without any lighting all over France and we obtain some observations  $X_i$ ,  $i = 1, \dots, 950$ , where  $X_i$  is a vector of length 16 giving the number of impacts at day  $i$  in each of the 16 regions. This numbers are highly non Gaussian, contain many zeros, and show a clear south-east/north-west tendency (with much more lightning in the south east). Therefore, we look at the numbers at the log scale (taking  $\log(X_i + 1)$ , with  $+1$  dealing with vanishing values  $X_i = 0$ ) and we subtracted the spatial tendency (this operation replaces vanishing values by small residues after regression). Now, it remains a strong inhomogeneity, that should violate the assumption that the underlying graph has no self-loops (*i.e.*, the diagonal of  $W$  is zero). To overcome this problem, we normalize the process in such manner that the conditional variance at each vertex conditionally to all the other is 1.

We model the resulting process as a spatial AutoRegressive gaussian fields, as described in Section 3.6. Given that the covariance matrix of the process commutes with the underlying graph, we applied our algorithm: we draw 100 learning samples, keeping or not each observation with probability 1/2, and retained 20% of the 100 trajectories. We do not obtain exactly the same graph at each run, although the graphs are most of the time very satisfying. To show the results, we ran 100 times the algorithm and kept, for every edge, the proportion of the time that this edge appears. This is summarized in Figure 7.

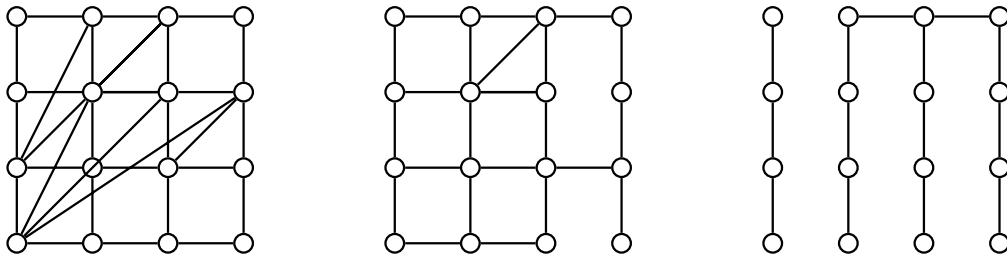


Figure 7: Edges that appears in 30%, 50%, and 70% of the time in the boosted backward algorithm.

To compare the performance of our method, we used the package **GGMselect** to infer Graphical Models, see [Giraud et al. \(2012\)](#). This package is very efficient, and powerful even for samples with more vertices than observations. It is not designed exactly for our case, so we do not pretend that our method makes better than this algorithm. Furthermore, we did not tune the parameters, and used rather the default parameters, only specifying the maximal degree of each vertex as `dmax = 5` and the family `C01`. The results are given in Figure 8.

We also implemented **GGMselect** on learning sample obtained keeping observation with probability 1/2 (as for the boosted backward algorithm), and represent how often an edge appeared, as in our method. We have to note that **GGMselect** seems more robust than our method, and this fact holds also for the other family `LA` even if we will not present here the quite similar results.

Furthermore, our algorithm takes a lot of time compared to `GGMselect` (0.3s. for `GGMselect` and 400s. for our algorithm.)

Nevertheless, the two methods give different results. The normalized lightning fields happens to be closed to a Simultaneous AutoRegressive process of order  $k$  on  $\mathbb{Z}^2$  (see for instance [Guyon \(1995\)](#) and [Gaetan et al. \(2010\)](#)). Hence, we expect that the target graph to be alike  $\mathbb{Z}^2$ . In Figure 7, we observe that the 50% present edges graph seems to uncover this spatial dependency keeping only edges between adjacent regions. Note that, the package `GGMselect` aims to recover a weighted graph of the paths of length at most  $k$  on the grid while our method aims to recover the graph itself.

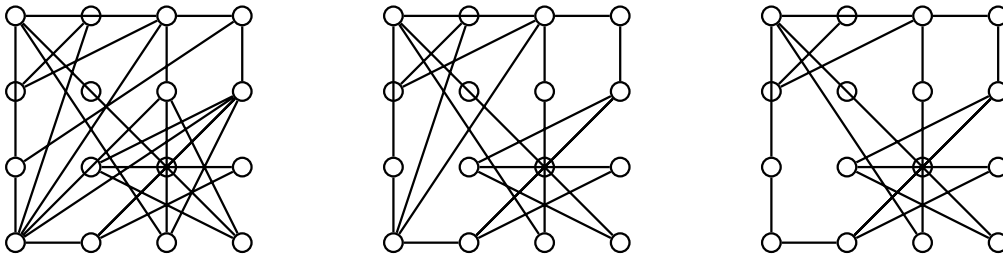


Figure 8: Results with the `GGMselect` package, with families LA, C01 and QE.

The results show that, in this case, and with the purpose of finding an underlying graph that “generates” the process, our method seems to work at least as fine as an inference of a graphical model, modeling data as a Gaussian Markov Field. We insist that we do not claim this fact to be general. In particular, we need much more observations than the methods developed in this package. But we pretend that, in different contexts, and with enough observations, we can be as good as other methods. Indeed, our method yet presents one advantage: the process does not need to be Markov, and for instance, we could infer spatial autoregressive process of any order (whereas graphical model inference can only recover underlying graphs for  $AR_1$  spatial processes, which are Markov). But this advantage turns into a problem when the process is truly Markov, because we do not use the knowledge of the function  $f$ , which can be taken as  $1/x$  in the Markov case.

## 7. Discussion

In this paper, we develop a new method to recover hidden graphical structures in different models that shares the fact that, one way or another, we have access to an approximation of the eigenstructure of the graph, through an estimation of an operator that commutes with a weighted adjacency matrix of this unknown graph. This is noticeable that we do not need any sparsity assumption to make the method work, and even with the large number of unknown parameters ( $K = f(W)$ , with the support, the function  $f$ , and the non-null entries of  $W$  are all unknown), we can perfectly recover the support when enough observations are available. We only assume that we know the location of some zeros. The most interesting case is when the known zeros are localized onto the diagonal, because it only means that the process is well normalized, in a sense, because all self-loops have same weights.

Note that there is a number of observations below which the algorithm always provides a wrong support. Furthermore, this fact can be observed in practice, because almost all learning samples will lead to different supports. This limit is intrinsic to our model and is a matter of balance between the sample noise  $\|\hat{K} - K\|$  and the signal strength. The noise is the estimation error of  $K$ , and has order  $1/\sqrt{n}$ , whereas the signal is of order  $c_0(S^*)$ , see (4) and (10).

Furthermore, the paper addresses the problem of exact support recovery, which is way harder than to provide an approximation of the support. The performances presented in this paper were computed with default parameters, but manual tuning seems to improve a little bit the results. In particular, drawing learning samples with probability  $\frac{1}{2}$  may cause overfitting, and for very large samples, we do not always get 100% exact support recovery. This problem can be easily bypassed by either decreasing the size of learning samples, or increasing the thresholds. In the present version, 3 parameters have been empirically chosen : the size of learning samples, the number of boosting trajectories, and the way we regroup the results of these boosting trajectories. One challenge for future work is to justify these choices with theoretical results.

For practical issues, it remains three other challenges that have to be bypassed. The first one concerns the assumption about the symmetry of  $W$ , that should be released for real practical interest. The second concerns the assumption that  $W$  has a null diagonal. It remains to find an effective way to normalize the process when this assumption does not hold (the normalization used in Section 6 assume an autoregressive structure). Finally, our algorithm is greedy when the size of the graphs increases, and for large graphs, it would be really interesting to find a way to compute a cheap version of the criterion, and to compute the significance of the variable.

## Acknowledgement

The authors would like to thank Dieter Mitsche for fruitful discussions. We would like to warmly thank Météo France et Météorage for providing us the data used in Section 6. We would like to thank the Universidad de la Habana (Cuba) and the Centro de Modelamiento Matemático (Chile) for their hospitality.

## Appendix A. Asserting the Diagonal Identifiability

### A.1 Necessary and sufficient conditions

In this section, we focus on the  $F$ -identifiability in the special case where the set of forbidden entries is the diagonal  $F_{\text{diag}} := \{(i, i) : i \in [1, N]\}$ . Recall that a support  $S$  is  $F_{\text{diag}}$ -identifiable, or simply diagonally identifiable (DI), if for almost every matrix  $A \in \mathcal{E}(S)$ ,

$$BA = AB, \text{diag}(B) = 0, B = B^\top \implies B = \lambda A$$

for some  $\lambda \in \mathbb{R}$ . In other words, a support  $S$  is diagonally identifiable if almost every symmetric matrix  $A$  with support in  $S$  is uniquely determined, up to scaling, by its eigenspaces among symmetric matrices with zero diagonal. In this section, we provide both sufficient and necessary conditions on a support  $S$  to ensure the  $F_{\text{diag}}$ -identifiability. For this, we consider a simple undirected graph  $G_S = ([1, N], S)$  on  $N$  vertices with edge set  $S$ .

**Definition 14 (Induced subgraph)** For  $V \subseteq [1, N]$ , the induced subgraph  $G_S(V) = (V, S(V))$  is the graph on  $V$  with edge set  $S(V) = S \cap V^2$ .

**Proposition 15** For all support  $S \subseteq [1, N]^2$ , the set of invertible matrices in  $\mathcal{E}(S)$  is either empty or a dense open subset of  $\mathcal{E}(S)$ .

The proof is straightforward when writing the determinant of  $A \in \mathcal{E}(S)$  as a polynomial in its entries. Observe that by this property, finding one invertible matrix  $A$  in  $\mathcal{E}(S)$  guarantees that almost every matrix in  $\mathcal{E}(S)$  is invertible. In this case, we say that the graph  $G_S$  is invertible. Similarly, we say that  $G_S$  is diagonally identifiable if  $S$  is diagonally identifiable.

**Theorem 16 (Conditions for  $F_{\text{diag}}$ -identifiability)** Let  $S \subseteq \overline{F}_{\text{diag}}$  and  $G_S = ([1, N], S)$ .

1. **Necessary condition:** If  $S$  is diagonally identifiable then there exists a sequence of subsets  $V_3, \dots, V_{N-1} \subset [1, N]$  such that  $|V_k| = k$  and  $G_S(V_k)$  is invertible for all  $k = 3, \dots, N-1$ .
2. **Sufficient condition:** If there exists a nested sequence  $V_3 \subset \dots \subset V_{N-1} \subset [1, N]$  with  $|V_k| = k$  such that  $G_S(V_k)$  is invertible for all  $k = 3, \dots, N-1$ , then  $S$  is diagonally identifiable.

The gap between the sufficient and necessary conditions lies essentially in the fact that the sequence  $V_3, \dots, V_{N-1}$  need to be nested for the sufficient condition.

**Proof** We proceed by contradiction. For the necessary condition, let  $k \geq 3$  be such that  $G_S(V_k)$  is not invertible, for all subset  $V_k \subset [1, N]$  of size  $k$ . For  $A \in \mathcal{E}(S)$ , denote by  $\psi_0(A), \psi_1(A), \dots, \psi_N(A)$  the coefficients of the characteristic polynomial

$$\det(zI - A) = \sum_{j=0}^N \psi_j(A) z^j, \quad z \in \mathbb{R}.$$

Consider the matrix  $M_k(A) := \sum_{j=0}^k \psi_j(A) A^j$ . By Eq. (14) in [Espinasse and Rochet \(2016\)](#), we see that the  $(i, i)$ -entry of  $M_k(A)$  equals the sum of all minors of size  $k$  that do not contain the vertex  $i$ . Thus, the condition that  $G_S(V_k)$  is not invertible for all subset  $V_k$  of size  $k$  implies that  $M_k(A)$  has zero diagonal. On the other hand, the non-zero entries of  $M_k(A)$  are degree  $k$  polynomials in the variables  $A_{ij}, (i, j) \in \text{Supp}(A)$ . Therefore, the equality  $M_k(A) = \lambda A$  for some  $\lambda \in \mathbb{R}$  occurs for at most a countable number of  $A \in \mathcal{E}(S)$ . Since  $M_k(A)$  commutes with  $A$ , we deduce that  $S$  is not diagonally identifiable.

For the sufficient condition, we will need the following lemma.

**Lemma 17** *If there exists a subset  $V' \subset [1, N]$  of size  $N - 1$  such that  $G_S(V')$  is both DI and invertible, then  $G_S$  is DI.*

**Proof** We may assume that  $V' = [1, N - 1]$  without loss of generality. Let  $M'$  denote a symmetric  $(N - 1) \times (N - 1)$  matrix indexed on  $V'$  that is both invertible and diagonally identifiable, i.e. for all non-zero matrix  $A' \neq \lambda M'$ ,

$$M'A' = A'M' \implies \text{diag}(A') \neq 0.$$

To prove that  $G_S$  is DI, it suffices to find a symmetric matrix  $M$  with support  $S$  that is diagonally identifiable. Consider  $M$  defined by

$$M = \begin{bmatrix} M' & 0 \\ 0 & 0 \end{bmatrix}.$$

Let  $A$  be a matrix with zero diagonal that commutes with  $M$  and write

$$A = \begin{bmatrix} A' & a \\ a^\top & 0 \end{bmatrix}$$

for some  $a \in \mathbb{R}^{N-1}$ , with  $\text{diag}(A') = 0$ . The condition  $MA = AM$  can be stated equivalently as

$$\begin{cases} M'A' = A'M' \\ M'a = 0 \end{cases}$$

Since  $M'$  is invertible by assumption,  $a = 0$  and the only matrix  $A$  with zero diagonal that commutes with  $M$  is the null matrix. Thus,  $M$  is diagonally identifiable. ■

We now go back to prove the sufficient condition in Theorem 16. Assume that  $G_S$  is not diagonally identifiable, then by Lemma 17, neither is  $G_S(V_{N-1})$ . By iterating the argument, we conclude that  $G_S(V_3)$  is not diagonally identifiable. However, the only invertible graph on three vertices is the triangle graph, which is diagonally identifiable, leading to a contradiction. ■

**Remark 18** *The proof of Theorem 16 combines the results of Lemma 2.1 in Barsotti et al. (2014) and Eq. (14) in Espinasse and Rochet (2016). The first one is of topological flavor proving that the set of identifiable matrices is either dense or empty in the set of matrices with prescribed support. The paper Barsotti et al. (2014) does not address condition on identifiability and Lemma 2.1 in Barsotti et al. (2014) is not an identifiability result. The second ingredient is Eq. (14) in Espinasse and Rochet (2016). Actually, the paper Espinasse and Rochet (2016) contains a key combinatorial computation on the adjugate matrix of weighted graphs and, we must confess, it has been motivated by addressing a combinatorial calculus in the proof of identifiability. It gives part of the present proof (it proves that  $M_k(A)$  has zero diagonal in the proof of the necessary condition) but it is far from being its essence. The proof of the sufficient condition does not involve this calculus and proving the necessary part requires other simple but non trivial steps.*

## A.2 Proof of Proposition 4

From Claim (ii) in Theorem 16 and considering the nested sequence  $V_{N-1} \supset \dots \supset V_3$  obtained by removing the last vertex on the tail of the kite at each step, we deduce a simple and tractable sufficient condition for a graph  $G_S$  to be diagonally identifiable, namely that  $G_S$  contains the kite graph as a vertex covering (possibly not induced) subgraph.

### A.3 Existence of kites

The condition on containing the kite graph  $\nabla_N$  as a subgraph is mild in the sense that it is satisfied in the dense regime  $\log n/n$  by random graphs, as depicted in the following proposition.

**Proposition 19** *The existence of kite graphs in the Erdős-Rényi model occurs as follows. For any  $\omega(N) \rightarrow \infty$  and for  $G_N \sim G(N, p_N)$ , if  $p_N \geq (1/N)(\log N + \log \log N + \omega(N))$  then  $\mathbb{P}\{G_N \text{ has a kite of length } N\}$  tends to 1 as  $N$  goes to infinity.*

The proof makes use of the existence of a hamiltonian cycle which is a standard result in Random Graph Theory, see Corollary 8.12 in Bollobás (1998) for instance. This results shows that in the regime  $(\log N + \log \log N)/N$  an Erdős-Rényi graph is diagonally identifiable.

**Proof** We now present the proof of this fact. Let  $\omega(n) \rightarrow \infty$  and set

$$\begin{aligned} p_1 &:= (1/n)(\log n + \log \log n + \omega(n)/2), \\ p_2 &:= \omega(n)/(2n). \end{aligned}$$

Let  $G^{(1)}$  and  $G^{(2)}$  be two independent Erdős-Rényi graphs such that

$$G_n^{(1)} \sim G(n, p_1) \quad \perp\!\!\!\perp \quad G_n^{(2)} \sim G(n, p_2).$$

As shown in Corollary 8.12 in Bollobás (1998) for instance,  $\mathbb{P}\{G_n^{(1)} \text{ is hamiltonian}\}$  tends to 1 as  $n$  goes to infinity. Given a hamiltonian cycle  $C_n$  of length  $n$  in  $G^{(1)}$  one can construct a kite of length  $n$  using edges of  $G^{(2)}$  to connect a pair of vertices at distance 2 on the cycle  $C_n$ . Invoke the independence of  $G^{(1)}$  and  $G^{(2)}$  to get that this latter probability is

$$\mathbb{P}\{\{k, k+2\} \text{ is an edge of } G^{(2)} \text{ for some } k\} = \mathbb{P}\{B(n, p_2) > 0\},$$

where  $B(n, p_2)$  denotes the binomial law. Using Poisson approximation one gets that this probability tends to 1 as  $n$  goes to infinity. We deduce that the probability that the graph  $G = G_n^{(1)} + G_n^{(2)}$  has at least a kite tends to 1. Observe that  $G$  is an Erdős-Rényi graph of size  $n$  and parameter  $p = p_1 + p_2 - p_1 p_2 \leq p_n$  which concludes the proof. ■

### A.4 Proof of Theorem 5

Combining Proposition 19 and Theorem 16, we deduce the first point. In view of the first point of Theorem 16, we see that it is sufficient to find two isolated vertices to prove non-identifiability. Indeed, in this case, the kernel of the adjacency matrix has co-dimension at least 2 showing that all sub-graphs of size  $N - 1$  are not invertible. Furthermore, one knows (see Theorem 3.1 in Bollobás (1998) for instance) that the event “there is at least two isolated points” has sharp threshold function  $\log n/n$ . It proves the second point.

## Appendix B. Support reconstruction

### B.1 Proof of Theorem 7

Define  $\mathcal{S}_1 := \{S \in \mathcal{S} : |S| \leq |S^*|, S \neq S^*\}$  and  $\mathcal{S}_2 := \{S \in \mathcal{S} : |S| > |S^*|\}$ , clearly it holds  $\mathcal{S} = \{S^*\} \cup \mathcal{S}_1 \cup \mathcal{S}_2$ . We want to control the terms  $\mathbb{P}\{\widehat{S} \in \mathcal{S}_1\}$  and  $\mathbb{P}\{\widehat{S} \in \mathcal{S}_2\}$  separately and conclude in view of

$$\mathbb{P}\{\widehat{S} \neq S^*\} = \mathbb{P}\{\widehat{S} \in \mathcal{S}_1\} + \mathbb{P}\{\widehat{S} \in \mathcal{S}_2\}.$$



Since the Frobenius norm is sub-multiplicative, it holds, for all  $A \in \mathcal{E}(\overline{F})$ ,

$$\|A(\widehat{K} - K) - (\widehat{K} - K)A\| \leq \|A(\widehat{K} - K)\|_2 + \|(\widehat{K} - K)A\| \leq 2\|A\|\|\widehat{K} - K\|.$$

Thus, the quantity  $\|A\widehat{K} - \widehat{K}A\|$  for  $A \in \mathcal{E}(\overline{F})$  can be bounded from below and above by

$$\|AK - KA\| - 2\|A\|\|\widehat{K} - K\| \leq \|A\widehat{K} - \widehat{K}A\| \leq \|AK - KA\| + 2\|A\|\|\widehat{K} - K\|. \quad (11)$$

To bound the term  $\mathbb{P}\{\widehat{S} \in \mathcal{S}_1\}$ , we use (11) to remark that for all  $S \in \mathcal{S}_1$ ,

$$Q(S) = \min_{A \in \mathcal{E}(S) \setminus \{0\}} \frac{\|A\widehat{K} - \widehat{K}A\|}{\|A\|} + \lambda_n |S| \geq \min_{A \in \mathcal{E}(S) \setminus \{0\}} \frac{\|AK - KA\|}{\|A\|} - 2\|\widehat{K} - K\|.$$

It follows

$$\min_{S \in \mathcal{S}_1} Q(S) \geq \min_{S \in \mathcal{S}_1} \min_{A \in \mathcal{E}(S) \setminus \{0\}} \frac{\|AK - KA\|}{\|A\|} - 2\|\widehat{K} - K\| = c_0(S^*) - 2\|\widehat{K} - K\|. \quad (12)$$

The constant  $c_0(S^*)$  is positive by  $F$ -identifiability of  $W$ . Moreover, observe that

$$Q(S^*) = \min_{A \in \mathcal{E}(S^*) \setminus \{0\}} \frac{\|A\widehat{K} - \widehat{K}A\|}{\|A\|} + \lambda_n |S^*| \leq \frac{\|W\widehat{K} - \widehat{K}W\|}{\|W\|} + \lambda_n |S^*| \leq 2\|\widehat{K} - K\| + \lambda_n |S^*|, \quad (13)$$

where we used both Eq. (11) and the fact that  $WK - KW = 0$ . Combining (12) and (13), we get

$$\mathbb{P}\{\widehat{S} \in \mathcal{S}_1\} \leq \mathbb{P}\left\{\min_{S \in \mathcal{S}_1} Q(S) \leq Q(S^*)\right\} \leq \mathbb{P}\left\{\|\widehat{K} - K\| \geq \frac{c_0(S^*) - \lambda_n |S^*|}{4}\right\}.$$

To control the term  $\mathbb{P}(\widehat{S} \in \mathcal{S}_2)$ , we use that  $\min_{S \in \mathcal{S}_2} Q(S) \geq \lambda_n \min_{S \in \mathcal{S}_2} |S| \geq \lambda_n (|S^*| + 1)$ . By Eq. (13), it follows

$$\begin{aligned} \mathbb{P}\{\widehat{S} \in \mathcal{S}_2\} &\leq \mathbb{P}\left\{\min_{S \in \mathcal{S}_2} Q(S) \leq Q(S^*)\right\} \\ &\leq \mathbb{P}\left\{\lambda_n (|S^*| + 1) \leq 2\|\widehat{K} - K\| + \lambda_n |S^*|\right\} \\ &= \mathbb{P}\left\{\|\widehat{K} - K\| \geq \frac{\lambda_n}{2}\right\}. \end{aligned}$$

The proof of Theorem 7 follows directly by (H<sub>2</sub>). The corollary is a direct consequence using Borel-Cantelli's Lemma.

## B.2 Proof of Theorem 10

Since  $\Delta(K)\Phi_S$  is of full rank, the value  $\widehat{\beta}_S = (\Delta(\widehat{K})\Phi_S)^\dagger \Delta(\widehat{K})a_0$  is the unique solution to Eq. (5) with probability tending to one asymptotically. Since the value of  $\widehat{\beta}_S$  does not depend on  $a_0 \in \mathcal{A}_S$ , one can take  $a_0 = w$  in view of  $S^* \subseteq S$ . We obtain

$$\widehat{\beta}_S = (\Delta(\widehat{K})\Phi_S)^\dagger \Delta(\widehat{K})w = -(\Delta(\widehat{K})\Phi_S)^\dagger \Delta(W)\widehat{k}.$$

The result follows from Slutsky's lemma, using that  $(\Delta(\widehat{K})\Phi_S)^\dagger$  converges in probability towards  $(\Delta(K)\Phi_S)^\dagger$  and

$$\sqrt{n}(\Delta(W)\widehat{k} - \Delta(W)k) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \Delta(W)\Sigma\Delta(W)^\top).$$

## References

- Flavia Barsotti, Yohann De Castro, Thibault Espinasse, and Paul Rochet. Estimating the transition matrix of a Markov chain observed at random times. *Statistics & Probability Letters*, 94:98–105, 2014.
- Flavia Barsotti, Anne Philippe, and Paul Rochet. Hypothesis testing for markovian models with random time observations. *Journal of Statistical Planning and Inference*, 173:87–98, 2016.
- José Bento and Morteza Ibrahimi. Support recovery for the drift coefficient of high-dimensional diffusions. *IEEE Transactions on Information Theory*, 60(7):4026–4049, 2014.
- José Bento, Morteza Ibrahimi, and Andrea Montanari. Learning networks of stochastic differential equations. In *Advances in Neural Information Processing Systems*, pages 172–180, 2010.
- Béla Bollobás. *Random graphs*. Springer, 1998.
- Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 2016.
- T Espinasse, F Gamboa, and J-M Loubes. Parametric estimation for gaussian fields indexed by graphs. *Probability Theory and Related Fields*, 159(1-2):117–155, 2014.
- Thibault Espinasse and Paul Rochet. Relations between connected and self-avoiding hikes in labelled complete digraphs. *Graphs and Combinatorics*, to appear, 2016.
- Carlo Gaetan, Xavier Guyon, and Kevin Bleakley. *Spatial statistics and modeling*, volume 81. Springer, 2010.
- Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. Graph selection with ggmselect. *Statistical applications in genetics and molecular biology*, 11(3), 2012.
- Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck’s inequality. *Probability Theory and Related Fields*, pages 1–25, 2015.
- Xavier Guyon. *Random fields on a network: modeling, statistics, and applications*. Springer Science & Business Media, 1995.
- Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *The Journal of Machine Learning Research*, 11:1709–1731, 2010.
- Olga Klopp, Alexandre B Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *arXiv preprint arXiv:1507.04118*, 2015.
- G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press Boston, 1990.
- Nicolas Verzelen. *Gaussian graphical models and Model selection*. PhD thesis, Université Paris Sud-Paris XI, 2008.
- Nicolas Verzelen, Ery Arias-Castro, et al. Community detection in sparse random networks. *The Annals of Applied Probability*, 25(6):3465–3510, 2015.